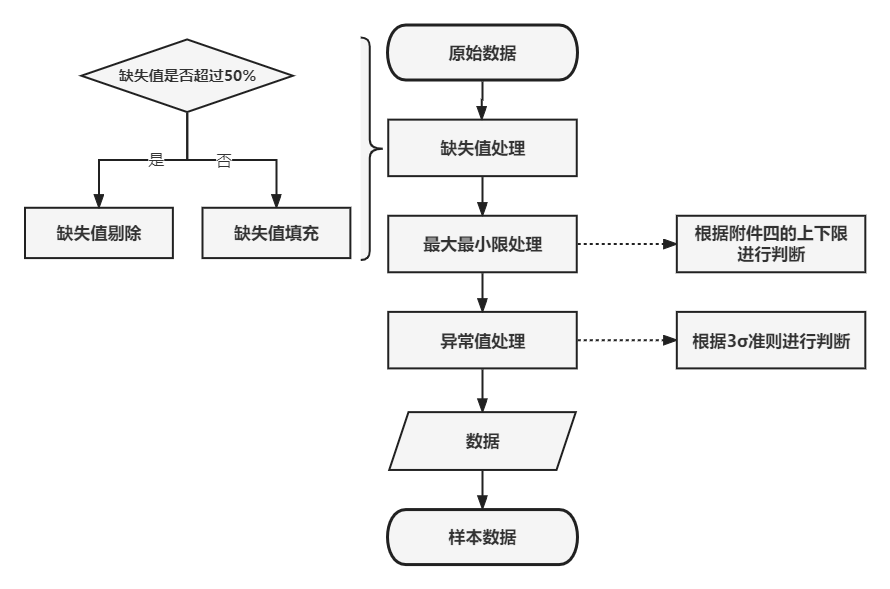
1. 问题一：样本及原始数据预处理
   1. 问题一分析

本文数据由两部分组成，分别为操作变量与非操作变量。其中操作变量来自于中石化高桥石化实时数据库（霍尼韦尔PHD），采集时间为2017年4月至2020年5月，采集了共**354个操作变量**。非操作变量来自LIMS实验数据库，采集时与操作变量一致，采集了共**13个非操作变量**。变量数据的采集频次与年份相关。而问题一的要求是根据附件二关于样本数据的处理方法，针对附件三的2个样本原始数据进行数据预处理，并加入到附件一的原始数据中，从而为解决后面的问题提供数据支撑。

对于实时数据库采集的操作变量数据，部分变量由于人为或非人为的原因会造成变量数据存在不完整、异常等显著误差，包括记录错误、操作错误、测量错误等，致使变量数据与实际数据无法始终保持一致，因此需要对这类变量进行数据预处理，从而为模型的建立提供干净的数据。而对于实验数据库采集的非操作变量，包括原料、待生吸附剂、再生吸附剂的相关变量，由于采集频次为每周2次，可以认为该类变量的动态变化范围较小，不需要进行数据预处理，且实验中假设该类变量的性质始终保持不变，因此问题一只需要对操作变量的数据进行处理。

在本文提供的原始数据中，大部分变量属于正常范围，但记录数据的装置在部分位点（变量）上存在问题，这会导致部分变量的数据在部分时间段为空值，或者部分样本的数据超出该变量的上下限以及存在异常。本文根据文献资料以及操作经验，针对样本中出现的问题，对这两个样本的数据通过以下流程进行处理：



* 1. 缺失值处理

如果某个变量仅含有部分时间点上的数据时，我们可以认为该变量存在缺失值，因此需要对其进行缺失值处理。缺失值处理一般有两种处理方式，分别为剔除和插补，其中插补的形式较多，包括均值插补、随即插补、回归插补和随机回归插补等，题目中指定用前后两小时的平均数进行插补，但在样本中仅提供了两个小时的样本数据，故本文选择用该变量在这两个小时的有效数据的平均值进行填充。

针对不同的情况需要采取不同的方式进行处理。在本文中，我们人为设定：当变量出现空缺的数量超过总样本数的50%时，需要对该变量进行剔除；当变量出现空缺的数量未超过总样本数的50%时，则可以对该变量进行插补。具体操作形式见下图：





其中样本数共有40个，a表示某个变量空缺值的个数，表示样本变量值。

我们分别对285样本和313样本进行缺失值处理，得到以下结果。

|  |  |
| --- | --- |
| 变量剔除数 | 变量插补数 |
| 11 | 0 |

|  |  |
| --- | --- |
| 变量剔除数 | 变量插补数 |
| 8 | 5 |

根据处理结果，可以发现285样本中仅出现了变量列全部为空值的情况，而313样本两种情况都有出现。

* 1. 最大最小限幅处理

根据附件4，可以得知各个操作变量的取值范围，因此通过python进行数据处理，从而得到样本变量是否有超过最大最小限的情况，如果存在，则对超出最大最小限的变量赋值为0，代表对该样本变量值进行剔除。

|  |  |
| --- | --- |
| 285样本数据剔除数 | 313样本数据剔除数 |
| 560 | 616 |

在进行这一步操作中，可以清楚发现，大量变量并不在变量的取值范围中，通过查看附件一的原始数据，仍然可以发现这些变量不仅在这两个样本中存在这种超限的异常情况，在其它样本中仍然存在，因此，这里我们推测部分异常情况是由于变量的单位未对应所造成，比如“%”，未将绝对数换算成相对数，从而导致变量数值上与最大最小限的数值存在较大的差异。

* 1. 异常值处理
     1. 达拉依准则

达拉依准则又称3准则，在统计学中较为流行，常用于判断数值是否出现较大的异常。在统计学中，我们认为某个数值的取值基本上集中于该数值平均数上下浮动3个单位的范围内。

在正态分布中，表示均值，表示标准差，3准则就是指，数值分布在的概率为99.74%，因此，我们可以认为数值的取值基本集中在这个区间范围，当出现数值不在这个区间的情况时，则认为该数值为异常值，应当予以剔除。

在本文中异常值处理主要是根据拉依达准则（3准则）来剔除，即取空值。



* + 1. 处理结果

通过异常值处理，我们在285样本中共发现了85个异常值，在313样本中共发现了181个异常值。

* 1. 数据处理结果和结论

通过缺失值处理、最大最小限幅处理与异常值处理，当前的数据已经符合样本数据的标准，我们对各样本的变量数据值取平均数，作为该样本这两个小时内的变量数据，以对应辛烷值的操作变量数据，并将其加入到附件一中。具体处理结果可以看附件三的

通过前后对比附件一285样本和313样本的数据变量，可以发现有几个变量值存在轻微浮动的情况，基本上只有小数点后几位的变动，由于计算上存在差异，可以认为这属于正常的浮动范围，因此我们认为这些变量数据能够进行后面几个问题的运算过程中去。